Guidelines for selecting the translation start site

Sections:

1. Definitions:

***Conservation*:**
We define conservation by observing sequence similarity for orthologous loci at the level of the genome sequence between two or more species with an emphasis (for curating human and mouse) on conservation observed in the genomes assemblies for human, chimp, macaque, mouse, rat, dog, and cow. Additionally, genome conservation may also be observed within a species for paralogous loci. Agreement with other independently curated datasets such as Swiss-Prot protein records may also be taken into consideration. Genome conservation may be observed using existing public tools, such as the UCSC Vertebrate conservation track, or in similar in-house tools provided to support curator staff.

a. Strong conservation: genome sequence is conserved in at least 2 species that are evolutionarily distant (e.g., different genus). Strong conservation support (or experimental data) is needed when considering a large N-terminal extension (>100aa).
b. Weak conservation: genome sequence is conserved in closely related species but not conserved in more distantly related species (e.g., such as within primates, within rodents, or within mouse strains)
c. ***Note***: Variation at the protein termini is valid and expected and can be lineage-specific. Small differences in N-terminal length between, for instance, human and mouse, are expected. Large differences may be valid but should be supported by available transcript, publication, and conservation data.

Kozak signal strength:

- GCC[A/G]CCaugG[not U] == optimal
- [A/G]NNaugG[not U] == strong; ‘A’ at -3 is stronger than ‘G’
- Anything else = ‘weak’

Some known modulators of initiation sites (general, not specific):

- Secondary structure:
  - A hairpin secondary structure downstream of a non-AUG initiation site, or downstream of an AUG with a weak Kozak signal, may facilitate pausing and increase the likelihood of initiation from that site (PMID: 2236042).
- uORF:
  - Short uORFs (< 35 aa) located upstream of the primary ORF is considered to have a role in reducing the translational efficiency of the primary ORF. The distance between the uORF and the primary ORF does not matter although there are some indications that a longer distance increases the efficiency of translation initiation at the primary ORF. The primary ORF is thus translated according to the leaky scanning model, or, for some genes, may be translated from different transcripts that do not include the uORF due to alternate promoter use or alternate splicing.
  - Long uORFs associated with a strong Kozak signal are considered to more severely inhibit translation; the definition of length is not absolute but Kozak proposes >35 as a general threshold. The theory is that leaky scanning does not come into play in the presence of a longer ORF with a strong Kozak signal. If the long uORF has a weak Kozak signal then the pORF may still be translated due to leaky scanning.
  - uORFs with a strong Kozak signal that *overlap* the primary functional ORF are thought to severely inhibit translation. Those with a weak Kozak may modulate translation but translation likely occurs from the downstream AUG due to leaky scanning.
- pORF: the primary ORF (pORF) considered to be the functional open reading frame based on homology, publications, and/or sequence content (protein domains etc).

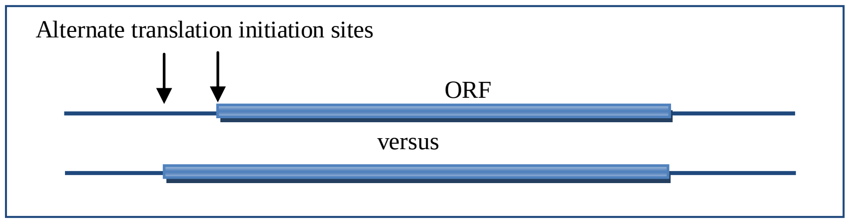
- This list is not intended to be comprehensive. Other factors do come into play and may vary in a gene-specific, developmental, or spatial manner.

Note: An upstream AUG with a strong Kozak signal should **always** be used irrespective of conservation, unless there is experimental evidence demonstrating otherwise. A strong Kozak signal would not be permissive to leaky scanning that would enable the use of a downstream start codon. The strength of the downstream Kozak signal is much less important (see PMIDs: 12459250 and 16213112).

Note: More support is required to annotate at an internal AUG site than at the first AUG site, when there are alternate possible start sites available on one transcript. This is because according to the scanning translation model, even if the first AUG is not in an optimal context then translation will generally still occur from that location at some frequency because the ribosome will pause at the AUG (e.g., the ribosome does not know *a priori* that a better site is available downstream). However, because the context is sub-optimal, the ribosome may not always pause long enough for initiation to *consistently* occur at the first AUG site, and a thus the ribosome may continue scanning, detect a downstream AUG, and initiate translation from that alternate site. If there are several AUGs in a weak context then initiation may actually occur at all sites. Once the ribosome encounters an AUG in a strong context then it will not continue scanning and initiation at any other downstream in-frame AUG sites could only occur by way of an alternate transcript that lacks the AUG with a strong Kozak signal (e.g., alternate promoter use or alternate splicing).

2. Curation Guidelines - start codon choice for CCDS:

2.A. Extension of the ORF – assessing alternate in-frame translation initiation sites



Default Rule: Always annotate the CDS starting from the upstream AUG ***unless*** one of the following exceptions applies.
Note: The expectation is that frequently none of the exceptions will apply; therefore, we will often annotate an upstream AUG with a weak Kozak signal and weak conservation.

1. Strong experimental evidence shows that the downstream start codon is used. Experimental evidence may include:

- Protein N-terminal sequencing.
- N-terminal-specific antibody support.
- Translation assays that include both start codons. Evidence showing that the downstream AUG can be used does not necessarily mean that the upstream start cannot be used.
- Evidence of 5’ UTR secondary structure/protein interactions that would preclude the use of the upstream AUG.
- Evidence that the shorter protein originating from the downstream AUG is the primary functional product (short protein is more abundant and function has been ascribed to the short protein).

2. The downstream AUG has a strong history of use and is considered to be the community standard. A ‘strong history’ may be indicated by several (5) high quality publications (or 10+ relevant publications) indicating consistent definition of the protein, especially the N-terminus or publications that cite point mutations at specific locations relative to a CDS standard. **Please contact a scientific expert** to determine if there is experimental support for either AUG in question, to determine if the community is aware of the upstream AUG, and to make a final decision regarding annotation.

Note: We expect that the upstream Kozak signal is not ‘strong’ and that genome conservation for the upstream AUG may also be weak.

Note: *there may be cases where new data must override historical use because historical use was based on incomplete knowledge*. A low number of publications perhaps based on an initial clone that is currently considered to be 5’ incomplete or that represents 5’ indels or that prevent the use of the upstream start codon, should not be sufficient evidence to annotate the downstream AUG.

3. The upstream AUG is not conserved in any other species AND the upstream AUG has a weak Kozak signal, which would permit leaky scanning:

- AND functional information is available for the shorter protein, either publication support or convincing domain structure that would be indicative of a function (e.g., always annotate the longest ORF for genes of unknown function)
- AND some combination of:
    - the N-terminal extension does not add a signal or transit peptide
    - the N-terminal extension does extend a signal peptide to an unreasonable length (roughly, >40 aa)***
    - the N-terminal extension is very large resulting in a protein length significantly different from either homologs or paralogs where the homolog/paralog is itself considered well supported and of full length and the difference is not due to alternate splicing. ***Note***: if the N-terminal extension under consideration is very large (say, ~>100aa) then require strong conservation support or experimental data.
    - the N-terminal extension does not add or complete a domain
    - the downstream AUG site has a strong Kozak signal context and/or is more strongly conserved

*** See distribution of eukaryotic signal peptide lengths at http://www.cbs.dtu.dk/services/SignalP-1.1/sp_lengths.html

Note: According to the scanning model, the ribosome does not know *a priori* whether the upstream or downstream AUG is conserved in other species.

4. A signal or transit peptide is predicted for the shorter, but not the longer, N-terminal AND the upstream AUG has a weak Kozak signal that would permit leaky scanning

- AND there is clear experimental evidence that leader peptide cleavage is necessary for a functional protein. E.g., there is experimental evidence in the literature indicating that this protein is secreted or targeted to a cellular compartment and the longer protein lacks the signal peptide.

5. There is an alternate start codon located 5’ of the first AUG (e.g., CTG, GTG, or ACG) that is:

- Experimentally verified as a translation initiation site (support could be from a homolog)
- Or, use of upstream non-AUG site completes a protein domain

- Or, completes or adds a signal or transit peptide (AND there is experimental support for a targeted location – support could be from a homolog)
- Or, there is very strong genome conservation for the alternate start site

Note: Kozak indicates that use of a non-AUG occurs if there is an optimal Kozak signal context and/or the mRNA structure supports pausing of the ribosome over the non-AUG site long enough to allow the inaccurate pairing of the initiation Met and the non-AUG codon. There may be leaky scanning and initiation at an alternate downstream AUG site in addition to initiation at the upstream non-AUG site. Therefore, we should have good support for a decision to annotate the CDS initiating at a non-AUG site.
Note: A hairpin secondary structure downstream of a weak Kozak site may facilitate ribosome pausing and thus increase the likelihood of initiation from that site (PMID: 2236042). There are hairpin prediction programs available, one is http://gibk26.bse.kyutech.ac.jp/aug_hairpin/

6. There is extremely strong conservation for a downstream AUG site, or any case that does not cleanly fit into the above guidelines:
   - Final annotation decision is made following discussion among the CCDS collaborators. This discussion may include consultation with other scientists.

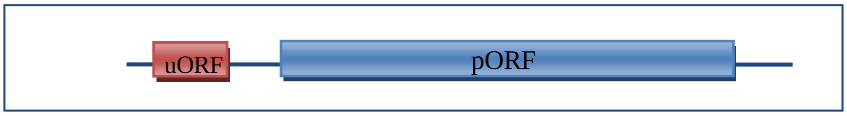
2.B. Upstream ORFs (uORF) vs. primary ORFs (pORF) – assessing translation probability

1. The transcript has a uORF that *overlaps* the pORF. (Kozak signal important, uORF length unimportant)

   a. If the uORF has a strong Kozak signal then the overlapping uORF strongly inhibits translation from the downstream AUG and you should assume that translation does initiate from the upstream AUG (the overlapping ORF).
      i. Annotate the uORF protein if it is supported by experimental evidence (e.g., it really does encode a protein that has been studied).
      ii. Annotate a non-coding transcript if the uORF is not supported by experimental evidence and/or would result in NMD.
      iii. *Exception*: Annotate the pORF protein if this is the only transcript form observed for a locus with abundant transcript support, and there is no question that the locus is protein coding. This should be tracked as a known annotation conflict (overlapping uORF with strong Kozak). This exception reflects the goal of providing annotation for known proteins whenever possible.

b. If the uORF has a weak Kozak signal then leaky scanning allows translation from the downstream AUG (pORF).
    i. Annotate the pORF protein.
    ii. This should be tracked as an annotation observation (overlapping uORF with weak Kozak).

2. The transcript has a uORF that does not overlap the pORF:



a. The uORF is <35 aa. (Kozak signal not important) Translation from the downstream AUG is permitted via re-initiation if the uORF Kozak is strong, and via leaky scanning or re-initiation if the uORF Kozak is weak; therefore, Kozak strength of uORF does not matter
    i. Always annotate the pORF protein.
    ii. Note: this case is very common

b. The uORF >=35 aa. (Kozak signal important)
Note: Consider if the uORF is a fragment of the CDS for the functional protein (as opposed to an unrelated uORF); use the uORF logic documented here to decide whether to annotate a protein from an internal AUG or to represent as a non-coding RNA and also track as a transcript that is also a NMD candidate if translation were to initiate at the upstream AUG.
    i. If the uORF has a strong Kozak signal then assume translation does initiate at first AUG. Neither leaky scanning nor re-initiation will occur and so translation will not initiate at the downstream AUG for the pORF.
        1. Annotate the protein for the uORF if there is published experimental support for translation of that open reading frame.
        2. If a protein for the pORF can be represented from a different transcript variant, then the transcript with the uORF and/or NMD conflict should be represented as a non-coding transcript unless there is direct experimental support for translation from this transcript form.
        3. *Exception:* Annotate the protein for the pORF if this is the only transcript form known for the locus, and the locus is unquestionably protein-coding. This should be tracked as a known annotation conflict (uORF >35aa with strong Kozak signal). This exception reflects the goal of providing annotation for known proteins whenever possible.
    ii. If the uORF has a weak Kozak signal then leaky scanning allows some translation from downstream AUG.
        1. Always annotate the pORF protein.
        2. This should be tracked as an annotation observation (uORF >=35 aa with weak Kozak)

3. *Example Cases:*

For annotating the CDS starting from the first AUG:

a) There is a strong Kozak signal for the first AUG and it is an extension of the primary ORF (regardless of the Kozak signal for the downstream AUG, and regardless of genome conservation).
b) There is a weak Kozak signal but there is strong genome conservation for the first AUG and the extension doesn't conflict with experimental information about translation initiation or localization. In this context, strong conservation at the level of genome sequence is observed between two or more species; the species do not need to be widely diverged (e.g. primate-specific N-terminal differences are valid).
c) There is a weak Kozak signal and there is weak or no conservation for the first AUG, but the extension improves the protein with regard to adding or completing a domain or signal/transit peptide.
d) There is no functional information, whether direct or indirect (domains), for the protein function.

For annotating the CDS starting from an internal AUG:

a) There is a weak Kozak signal and no conservation for the first AUG, and very strong conservation and a strong Kozak signal at a downstream AUG. There is significant genome conservation observed among species with evolutionary distance and there is consistency in the location of the downstream AUG site and N-terminus region of the protein.
b) There is a very strong historical use; the protein as defined from the internal AUG is considered the community reference standard. ***Note***: if you think this is a case where newer data indicates historical use is faulty then it may be useful to consult with an expert on the gene/protein to confirm the N-terminus representation. The community standard N-terminus should be supported by available public data. In other words, there should be a compelling reason to not annotate from the upstream AUG especially when there is conservation support or a good Kozak signal. In one real case, the community expert pointed out that the upstream AUG site being considered was invalid because the transcript representation was in error; promoter studies determined that the predominant transcript start occurred after the first AUG site that was in question. The transcript representation had been extended further 5' of the known promoter based on weak transcript support that the scientific expert did not consider valid.
c) ***Note***: if the 'internal' AUG site in question is also available as the first AUG site on a different transcript due to use of an alternate promoter, or alternate splicing, then (given sufficient

support) both transcripts and both N-terminal protein options can be annotated. Naturally, all transcripts have to themselves meet quality and abundance criteria to be considered representing as annotated alternate transcripts.

Cases where a leader peptide can be predicted for both N-termini are less clear and may require further discussion, with consideration for the signal peptide length.

4. References:

1. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes.(2005) Gene. 361:13-37. (PMID: 16213112)
2. Kozak M. Pushing the limits of the scanning mechanism for initiation of translation. (2002) Gene. 299(1-2):1-34. (PMID: 12459250)
3. Kozak M. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. (1990) Proc Natl Acad Sci 87(21):8301-5. (PMID: 2236042)